MANPOWER PLANNING MODELS. IV. SYNTHESIS OF CROSS-SECTIONAL AND LONGITUDINAL MODELS

R. C. Grinold, et al

Naval Postgraduate School
Monterey, California

November 1975

NPS55Mt75111

NAVAL POSTGRADUATE SCHOOL

Monterey, California

MANPOWER PLANNING MODELS - IV

SYNTHESIS OF CROSS-SECTIONAL AND LONGITUDINAL MODELS

by

R. C. Grinold

and

K. T. Marshall

November 1975

Approved for public release; distribution unlimited.

Prepared for·
Navy Personnel R&D Center
San Diego, CA 92152

NAVAL POSTGRADUATE SCHOOL
Monterey, California

Rear Admiral Isham Linder
Superintendent

Jack R. Borsting
Provost

This work was supported in part by the Navy Personnel R&D Center, San Diego, Office of Naval Research and the Manpower Planning Division (MPI20) of the Marine Corps.

Reproduction of all or part of this report is authorized.

Prepared by:

Kneale T. Marshall
Kneale T. Marshall

R.C. Grinold (by KTM)
R. C. Grinold

Reviewed by:

David A. Schrady
David A. Schrady, Chairman
Department of Operations Research
and Administrative Sciences

Released by:

Robert Fossum
Robert Fossum
Dean of Research

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NPS55Mt75111	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Manpower Planning Models - IV Synthesis of Cross-Sectional and Longitudinal Models		5. TYPE OF REPORT & PERIOD COVERED Technical Report
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) R. C. Grinold K. T. Marshall		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Postgraduate School Monterey, CA 93940		10. PROGRAM ELEMENT PROJECT, TASK AREA & WORK UNIT NUMBERS N6822176WR60008
11. CONTROLLING OFFICE NAME AND ADDRESS Navy Personnel R&D Center San Diego, CA 92152		12. REPORT DATE November 1975
		13. NUMBER OF PAGES 43
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS (of this report) Unclassified
		15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Manpower Planning Models Flow

Markov Chains

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report is the fourth in a series on Manpower Planning Models. Its main purpose is to compare the cross-sectional and longitudinal models described in the second and third reports, point out their similarities and differences, and present a theoretical comparison of the two types of models.

TABLE OF CONTENTS

IV. SYNTHESIS OF CROSS-SECTIONAL AND LONGITUDINAL MODELS

1. Introduction.

This chapter examines the relationships between the cross-sectional models developed in Chapter II and the longitudinal models developed in Chapter III. The longitudinal models allow more general flow processes to be modelled, and any cross-sectional model is a special case of a longitudinal model. Although the longitudinal models are more general, they normally have much greater data requirements and thus are more difficult to implement in cases where the model coefficients are estimated from historical data. Therefore we seek some compromise between the basic longitudinal and cross sectional models.

The chapter begins with a brief section demonstrating some relationships between the two models. Sections 3 and 4 present hybrid models that use cross-sectional data yet have some longitudinal characteristics. Section 3 describes two characteristic models. These large cross-sectional models have a special structure which allows for simple calculations and modest data requirements. Section 4 considers semi-Markov models which are a straight forward extension of the cross-sectional model. We find that the special structure of the semi-Markov model yields some useful approximations. Finally, section 5 is devoted to a theoretical analysis of the longitudinal model and the analysis of errors caused by using a best approximating cross-sectional model.

In this chapter we modify our previous notational conventions. When it simplifies the exposition we assume that the longitudinal matrices $P(u)$ will have index u for all u greater than or equal to zero. In previous chapters we assumed that $P(u) = 0$ for $u > M$. This case is still included of course, but allowing u to range over all positive values often simplifies the limits on summations in complicated expressions. We also use the probabilistic interpretations of the cross-sectional and longitudinal models. With the exception of section 5 all the arguments could be reworded in terms of fractional flows. However, the use of the probabilistic nomenclature eases the discussion and simplifies some of the arguments.

2. Relations Between Cross-Sectional and Longitudinal Models.

This section contains an analysis of the relations between cross-sectional and longitudinal models. It starts with the introduction of an expanded classification scheme which connects the two models. This leads us to examine several practical considerations in class expansion. A detailed theoretical analysis of model comparison is given later in section 5.

In order to use the cross-sectional models described in Chapter II one must first select a suitable manpower classification scheme. In general one selects the simplest scheme that will answer specific interesting questions, and stay consistent with available data. It may be helpful to expand the classification scheme to develop a more realistic model of the flow process.

The cross-sectional data found in most organizations often contains limited longitudinal information. For example, in a faculty promotion model such as that described in II.8, the data on individual faculty members probably contains, in addition to current rank, the length of time in the organization, or length of time in the current rank. This data often indicates how a simple classification scheme, such as rank, can be expanded to more realistically model personnel flows. We exploit this idea below, but first we see how a general longitudinal model can be rearranged and thought of as a cross sectional model.

Recall from the general longitudinal model in III.2 that the input flows on chains 1 through K in period t are given by the K-vector $g(t)$, and the maximum number of periods spent in the system is $M + 1$. Suppose that we define a class to be a combination of chain-type and period of entry. Then we have $K \times (M + 1)$ classes. Let the "stocks" at time t be given by the $K \times (M + 1)$-vector of past chain input flows $[g(t), g(t-1),\ldots,g(t-M)]$, and Q be a

$K \times (M + 1)$ square matrix with zeros except for 1's on the K-th lower diagonal. If $\bar{0}$ represents a $K \times K$ zero matrix, and I a $K \times K$ identity matrix, then for $M = 3$,

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \end{bmatrix} .$$

Let $f(t)$ be a $K \cdot (M + 1)$-vector whose first K elements are $g(t)$ and the remainder all zeros. Then

$$s(t + 1) = Qs(t) + f(t)$$

and we have a cross-sectional formulation. However, the model is simply a reorganization of the general longitudinal model. We now look at some particular cases of more interest.

Suppose $P(0)$ is a given $(N \times K)$ matrix and $P(u + 1) = QP(u)$, where Q is an $N \times N$ matrix. Then, for all u, $P(u + 1) = Q^{u+1}P(0)$, and using equation (4) in III.2,

$$s(t) = P(0 g(t) + Q \sum_{u=1}^{\infty} Q^{u-1}P(0)g(t - u)$$

(1)

$$= Qs(t - 1) + P(0)g(t) .$$

This is a cross-sectional model with $f(t) = P(0)g(t)$.

A converse to this result is also true. Suppose $s(t) - P(0)g(t) = Qs(t - 1)$ for any values of $g(t - u)$, $u \geq 1$. Then we must have $P(u + 1) = Q^{u+1}P(0)$. To see this set $g(t - u) = 0$, except when $u = k$. Then $s(t - k) = P(0)g(t - k)$ and $s(t) = P(k)g(t - k) = Q^{k}P(0)g(t - k)$. Since $g(t - k)$ is arbitrary, we must have $P(k) = Q^{k}P(0)$. Thus we have shown the longitudinal and cross-sectional models are identical if and only if $f(t) = P(0)g(t)$ and $P(u + 1) = Q^{u+1}P(0)$ for all $u \geq 0$.

Problem 1: If $P(u + 1) = Q^{u+1}P(0)$, and the maximum number of periods in the system is $M + 1$, what limitations does this place on the structure of Q?

Returning to the expansion of the classification scheme suppose that we have a longitudinal model with N classes, and maximum time in system equal to $(M + 1)$ periods. A class is now redefined to be a combination of an original class i and a length of completed service u. Thus there are $N \times (M + 1)$ new classes, and the stocks in these classes are given by the vector $[s_i(t;u)]$, for $i = 1,2,\ldots,N$, and $u = 0,1,2,\ldots,M$. Consider first the special case where the number of original classes N is equal to the number of chains K. Thus the matrices $P(u)$ in the longitudinal model are each square.

Define $q_{ji}(u)$ as the fraction of those in original class i with u periods of completed service, who move to original class j in one period. Then for each $k = 1,2,\ldots,K$,

$$p_{jk}(u + 1) = \sum_{i=1}^{N} q_{ji}(u)p_{ik}(u) ,$$

or

$$P(u + 1) = Q(u)P(u) .$$

If $P(u)$ has an inverse, then

$$Q(u) = P(u + 1)P(u)^{-1} \quad \text{for} \quad u = 0,1,\ldots,M - 1 .$$

In this case, the cross-sectional model is

$$s(t + 1;0) = g(t + 1) ,$$

$$s(t + 1; u + 1) = Q(u)s(t;u) \qquad u = 0,1,\ldots,M - 1 .$$

Example 1:

In the one class one chain model $(K = N = 1)$ we have $q(u) = p(u + 1)/p(u)$. If $p(0) = 1$, and $p(u)$ is nonincreasing, then $0 \leq q(u) \leq 1$. The numbers

$q(u)$ are commonly called *continuation rates,* since $q(u)$ gives the fraction of people who continue in the system for at least $(u + 1)$ periods, given that they have been in the system u periods. ■

More generally, when $N \neq K$, we can choose $Q(u)$ so that $Q(u)P(u)$ approximates $P(u + 1)$. This can be accomplished if, for each $j = 1,2,\ldots,N$, we solve the quadratic minimization problem:

$$\text{Minimize} \sum_{k=1}^{K} v_k^2$$

where

$$v_k = \sum_{i=1}^{N} q_{ji}(u)p_{ik}(u) - p_{jk}(u + 1) .$$

The matrix $Q(u)$ which solves this problem is given by

$$Q(u) = P(u + 1)P(u)^+ ,$$

where $P(u)^+$ is the generalized inverse of $P(u)$. However, there is no guarantee $Q(u)$ will be nonnegative with column sums less than one.

We close this section with a practical discussion of how a model with longitudinal features can be modified to seem more like a cross-sectional model. It seems best to establish this point by example.

Example 2: Consider the three class cross-sectional faculty model in example 1 of II.3. Given an individual enters class 1, the individual can move eventually to class 0 or 2. The expected duration in class 1 is $\frac{1}{1-q_{11}}$. If we ask for the expected duration conditioned on moving to class 0 (is not given tenure) the answer is still $\frac{1}{1-q_{11}}$. The same answer will be obtained if we ask for the expected lifetime in class 1 given eventual promotion to class 2 (is given tenure). The Markov model treats a visit to class 1 as a two-stage process, as is illustrated in Figure IV.1.

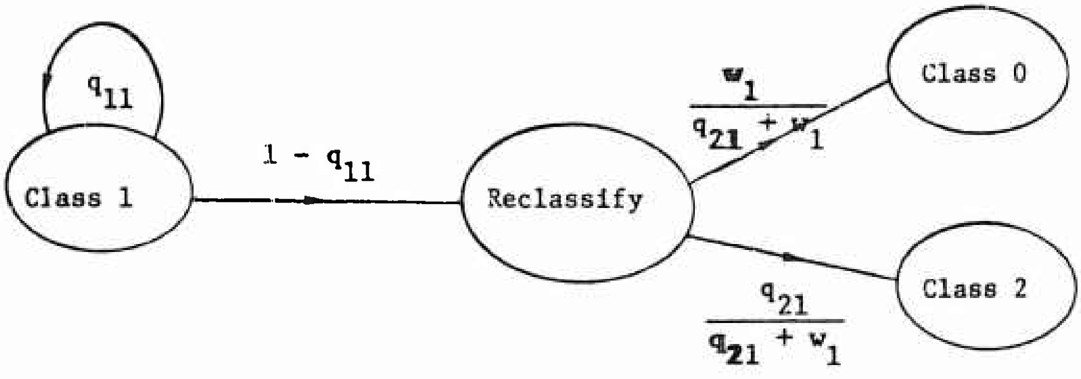

Figure IV.1. Illustration of Markov Model in Example 2.

At the first node, the individual either stays in class 1 or not and the expected number of periods at class 1 is independent of the reclassification process.

Suppose we know that the lifetimes of individuals in class 1 are dependent on their eventual status. Let T_0 be the expected lifetime in class 1 given an eventual move to class 0, and T_2 be the expected lifetime in class 1 given an eventual move to class 2. We can construct a four class cross-sectional model that has these characteristics:

New Class	Old Class
1. Nontenure who leave	1. Nontenure
2. Nontenure who move to tenure	
3. Tenure	2. Tenure
4. Retired	3. Retired

The new system will be distinguished by a ~

$$\tilde{s}(t) = \tilde{Q}\tilde{s}(t - 1) + \tilde{f}(t) .$$

We assume that

$$\tilde{f}_1(t) = \frac{w_1}{w_1+q_{21}} f_1(t) \ ,$$

$$f_2(t) = \frac{q_{21}}{w_1+q_{21}} f_1(t) \ ,$$

and

$$\tilde{Q} = \begin{bmatrix} \frac{T_0-1}{T_0} & 0 & 0 & 0 \\ 0 & \frac{T_2-1}{T_2} & \frac{1}{T_2} & 0 \\ 0 & 0 & q_{22} & q_{23} \\ 0 & 0 & 0 & q_{23} \end{bmatrix} .$$

This expanded model makes the distinction we desire in time spent in nontenure, and it also tells us the fraction of professors in nontenure that eventually acquire tenure, namely $\tilde{s}_2(t)/(\tilde{s}_1(t) + \tilde{s}_2(t))$.

■

3. Two-Characteristic Cross-Sectional Models.

This section examines cross-sectional models with two dimensional state spaces using the probabilistic interpretation presented in III.9. Assumptions on permissible flows between states lead to a special structure, and this in turn allows simple calculation of quantities such as projected inventories and lifetime in each classification.

The key to the special structure is the organization of the classification scheme. The classes (or states) are defined in terms of two characteristics, (i,j), where the first characteristic (henceforth FC), i, runs over the indices 1 through N. The range of the second characteristic (henceforth SC), j, depends on the FC. Let $\mathcal{S}$ be the set of all possible classes, and $\mathcal{S}(i) = \{j \mid (i,j)\varepsilon\mathcal{S}\}$ be the set of possible SC's given that the FC is i. Let $|\mathcal{S}(i)|$ be the number of elements in the set $\mathcal{S}(i)$.

At time t an individual's class can be described by a random variable $X(t)$. The cross-sectional assumption assures us that knowledge of $X(t)$ is sufficient for prediction of $X(t + 1)$, $X(t + 2)$, etc., without knowledge of $X(t - 1)$, $X(t - 2)$, etc. To obtain the special structure of the two characteristic model we impose limitations of the allowable transitions between classes. If the current FC is i, the only allowable moves in one period are

(i) to classes with FC still equal to i,

or (ii) to classes with FC equal to $i + 1$.

Example 3: Let the FC represent length of time in system and SC the grade of an individual. Consider the four grade student example with grades $j = 1,2,3,4$, for freshman, sophomore, junior and senior respectively. Clearly in each time period the first characteristic increases by 1. Let the maximum time in the system be 5 years (1 year = 1 time period), and let the sets of classes be

i	$\mathcal{S}(i)$
1	$\{1\}$
2	$\{1,2\}$
3	$\{2,3\}$
4	$\{3,4\}$
5	$\{4,5\}$

This is an example of the 'LOS/GRADE' model. Note that $N = 5$, and $|\mathcal{S}| = 9$.

<u>Problem 2</u>: List all the chains which would be present if example 3 were re-formulated as a longitudinal model.

Since the two-characteristic model is of the cross-sectional type it must be defined by a transition matrix Q, where Q is square with each dimension equal to $|\mathcal{S}|$. We consider the two types of allowable flow separately.

(i) No change in FC i.

Define for each j and m in $\mathcal{S}(i)$,

$$q_{mj}(i) = P[X(t + 1) = (i,m)|X(t) = (i,j)] ,$$

and let $Q(i)$ be the $|\mathcal{S}(i)|$ by $|\mathcal{S}(i)|$ matrix with (m,j)-th element equal to $q_{m,j}(i)$.

(ii) Change from FC i to FC $(i + 1)$.

Define for each m in $\mathcal{S}(i + 1)$ and each j in $\mathcal{S}(i)$,

$$p_{mj}(i) = P[X(t + 1) = (i + 1,m)|X(t) = (i,j)] ,$$

and let $P(i)$ be the $|\mathcal{S}(i + 1)| \times |\mathcal{S}(i)|$ matrix with (m,j)-th element equal to $p_{mj}(i)$.

The Q matrix is given by (for N = 4)

$$Q = \begin{bmatrix} Q(1) & 0 & 0 & 0 \\ P(1) & Q(2) & 0 & 0 \\ 0 & P(2) & Q(3) & 0 \\ 0 & 0 & P(3) & Q(4) \end{bmatrix}, \tag{2}$$

where the 0's are matrices with all elements equal to zero.

Example 4: Continuation of example 3.

Since the LOS must increase by 1 each year all the Q(i) matrices are zero matrices. Thus Q has the structure

$$Q = \left[\begin{array}{c|cc|cc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & x & x & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & x & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & x & x & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & x & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & x & x & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

where x indicates a (possibly) non-zero element. The partitioning is included to help the reader identify the P(i) matrices.

Example 5: Re-formulation of example 3.

Suppose that the FC represents the grade of an individual in a system where no demotions can occur and in which a person cannot advance more than one grade per year. Let SC represent the time spent in the particular grade. This is called the 'GRADE/TIME-IN-GRADE' model. Let the grades be 1) freshman, 2) sophomore,

3) junior and 4) senior, and let the maximum time in each grade be 2 years. Thus we have

i	S(i)
1	{1,2}
2	{1,2}
3	{1,2}
4	{1,2}

Note that $N = 4$ and $|S| = 8$. Now the Q matrix has the structure.

$$Q = \left[\begin{array}{cc|cc|cc|cc}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
x & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
x & x & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & x & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & x & x & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & x & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & x & x & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & x & 0
\end{array}\right]$$

where again x indicates a (possibly) non-zero element.

Example 6: Re-formulation of example 3.

Suppose that the FC represents the grade of an individual (as in example 5) in a system with no demotions and no double or multiple promotions per period. Let the SC represent the time in the system, or length of service (LOS). This is called the 'GRADE/LOS' model. Let the grades be 1) freshman, 2) sophomore, 3) junior and 4) senior, and let the maximum time in the system be 5 years, with

i	$\mathcal{S}$(i)
1	{1,2}
2	{2,3}
3	{3,4}
4	{4,5}

Note that N = 4 and $|\mathcal{S}| = 8$. Now the Q matrix has the structure

$$Q = \left[\begin{array}{cc|cc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ x & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline x & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & x & x & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & x & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & x & x & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & x & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & x & x & 0 \end{array}\right].$$

■

All the above examples display the special structure of Q which is depicted in (2). Recall from Chapter II that many applications of the cross-sectional model require calculation of the inverse $(I-Q)^{-1}$ which we called D. Although the Q matrix in the two-characteristic model is often quite large, it is easy to calculate D in terms of the inverses of the smaller submatrices. Define $D(i) = (I-Q(i))^{-1}$ for each FC i. Then (for the case N = 4),

$$D = \begin{bmatrix} D(1) & 0 & 0 & 0 \\ D(2)P(1)D(1) & D(2) & 0 & 0 \\ D(3)P(2)D(2)P(1)D(1) & D(3)P(2)D(2) & D(3) & 0 \\ D(4)P(3)D(3)P(2)D(2)P(1)D(1) & D(4)P(3)D(3)P(2)D(2) & D(4)P(3)D(3) & D(4) \end{bmatrix}.$$

Thus D is completely determined by the matrices $D(i)$, $i = 1,\ldots,N$, and $P(i)$, $i = 1,2,\ldots,N - 1$.

Computations in forecasting are considerably reduced by taking advantage of the special structure. Let $s_i(t)$ be the vector of stocks at time t with FC i. Thus $s_i(t)$ is a $|\mathcal{S}(i)|$ vector. Then the stocks at $(t + 1)$ are given by

$$s_i(t + 1) = Q(i)s_i(t) + P(i-1)s_{i-1}(t) + f_{0i}(t + 1), \quad i = 2,\ldots N ,$$

where $f_{0i}(t)$ is the vector of input flows in period t with FC i. The total stocks at $(t + 1)$ with FC i is found by summing the elements of $s_i(t + 1)$.

Problem 3: Let $b_{mj}(i)$ be the probability that, given the current state is (i,j), the state entered on leaving $\mathcal{S}(i)$ is $(i + 1,m)$. Let $B(i) = [b_{mj}(i)]$. Show that $B(i) = P(i)D(i)$.

Problem 4: Let $b_{mj}(k;i)$ be the probability that, given the current state is (i,j), the state entered when $\mathcal{S}(k)$ is entered is (k,m). Let $B(k;i) = [b_{mj}(k;i)]$, an $|\mathcal{S}(k)|$ by $|\mathcal{S}(i)|$ matrix. Show that $B(i) = B(i + 1;i)$, and that for $k > i+1$ $B(k;i) = B(k - 1)B(k-2) \ldots B(i)$.

■

4. Semi-Markov Flow Models.

A simple longitudinal model that retains some of a cross-sectional model's useful properties is the semi-Markov model. This section presents the general ideas behind such a model and indicates how some useful quantities can be calculated or approximated without completely specifying the flow process. We use terminology from probability theory to present the model, but the reader should recall that it is not necessary to view the model in a probabilistic sense. Although it can be viewed as a deterministic flow process we find the exposition easier and smoother using Markov chain terminology.

Consider a system with N classes of manpower. When an individual enters class i we say he commences a visit to class i. Let $q_{ji}(u)$ be the probability that a visit to class i lasts u periods and finishes with transition to state j. As in earlier chapters class 0 is interpreted as outside the system, and since a visit to any class is assumed to be at least 1 period in length, $q_{ji}(0) = 0$.

The probabilities $q_{ji}(u)$, $i = 1,2,\ldots,N$, $j = 0,1,2,\ldots,N$, $u = 1,2,\ldots,$ form the basic data of the model, and from these the following interesting quantities can be calculated:

(i) the probability that class j will follow class i,

$$q_{ji} = \sum_{u=1}^{\infty} q_{ji}(u) \ ,$$

(ii) the expected length of a visit to class i, given j is the next class visited,

$$\mu_{ji} = \sum_{u=1}^{\infty} uq_{ji}(u)/q_{ji} \ ,$$

(iii) the expected length of a visit to class i,

$$\mu_i = \sum_{j=0}^{N} \mu_{ji}q_{ji} \ ,$$

(iv) the probability of spending more than u periods in class i,

$$h_i(u) = \sum_{v=u+1}^{\infty} \sum_{j=0}^{N} q_{ji}(v) ,$$

(v) the variance in the length of a visit to class i, given that the next class visited is j,

$$\sigma_{ji}^2 = \sum_{u=1}^{\infty} (u - \mu_{ji})^2 q_{ji}(u)/q_{ji} ,$$

(vi) the variance in the length of a visit to class i,

$$\sigma_i^2 = \sum_{u=1}^{\infty} \sum_{j=0}^{N} (u - \mu_i)^2 q_{ji}(u) .$$

Problem 5: Show that

$$\mu_i = \sum_{u=0}^{\infty} h_i(u) ,$$

and

$$\sigma_i^2 + \mu_i^2 - \mu_i = 2 \sum_{u=0}^{\infty} u h_i(u) .$$

Example 7: Consider a student enrollment model with the following 5 states:

1. Freshman
2. Sophomore
3. Juniors
4. Seniors
5. Degree winners (graduates).

Assume that the only transitions possible are from i to either $(i + 1)$ or 0, and that no state can be held for more than three periods. The basic data are given by (blanks indicate zeros):

		u	
	1	2	3
$q_{01}(u)$	0.15	0.10	
$q_{21}(u)$	0.65	0.10	
$q_{02}(u)$	0.10	0.05	0.01
$q_{32}(u)$	0.70	0.10	0.04
$q_{03}(u)$	0.15	0.05	
$q_{43}(u)$	0.75	0.05	
$q_{04}(u)$	0.05		
$q_{54}(u)$	0.90	0.05	
$q_{05}(u)$	1.00		

By using (i) it is easy to calculate the 6 × 5 matrix of probabilities $[q_{ji}]$. These are:

j \ i	1	2	3	4	5
0	0.25	0.16	0.20	0.05	1.00
1					
2	0.75				
3		0.84			
4			0.80		
5				0.95	

Notice that the elements in each column sum to 1.00.

By using (ii) the expected values $[\mu_{ji}]$ are

j \ i	1	2	3	4	5
0	1.40	1.44	1.25	1.00	1.00
1					
2	1.13				
3		1.21			
4			1.06		
5				1.05	

From this table we see that, given a student will become a junior, the expected time he spends as a sophomore is 1.21 periods. Given he is to leave after being a sophomore, the expected time spent as a sophomore is 1.44 periods.

By using (v) the variances $[\sigma^2_{ji}]$ are

j \ i	1	2	3	4	5
0	0.24	0.37	0.19		
1					
2	0.12				
3		0.26			
4			0.06		
5				0.05	

The semi-Markov model can be viewed as a cross-sectional model with a two-characteristic state space (the reader should verify that the converse is not true). Suppose that a new state is defined to be a combination of an original state i and the number of periods spent in that state, u. Then an individual in state (i,u) moves next either to state $(j,0)$, with probability $q_{ji}(u + 1)/h_i(u)$, or to state $(i,u + 1)$ (remains in the same "original state") with probability $h_i(u + 1)/h_i(u)$.

Example 8: Continuation of example 7.

In this student example there are 10 states with a cross-sectional model Q matrix given by

	From									
To	(1,0)	(1,1)	(2,0)	(2,1)	(2,2)	(3,0)	(3,1)	(4,0)	(4,1)	(5,0)
(1,0)										
(1,1)	0.20									
(2,0)	0.65	0.50								
(2,1)			0.20							
(2,2)				0.25						
(3,0)			0.70	0.50	0.80					
(3,1)						0.10				
(4,0)						0.75	0.50			
(4,1)								0.05		
(5,0)								0.90	1.00	

Problem 6: In terms of the GRADE/TIME-IN-GRADE model described in Section 3, partition the matrix in example 8 to find the $Q(i)$ and $P(i)$ matrices, and find the inverse matrix $D = (I-Q)^{-1}$. Interpret the result.

■

The semi-Markov model can also be viewed as a longitudinal model, but in order to do this we must identify the chains. Chain k in the longitudinal interpretation corresponds to state k in the semi-Markov formulation. An individual is appointed in chain k if and only if he enters the system in state k. Recall from III.5 that $p_{ik}(u)$ is the probability that an individual who enters on chain k in some period t will be in class i at time $t + u$.

By using conditional probability arguments, when k is different from i we obtain from the semi-Markov assumptions,

$$\begin{aligned} p_{ik}(u) &= 0 && \text{if} \quad u = 0 \, , \\ &= \sum_{v=1}^{u} \sum_{j=1}^{N} p_{ij}(u-v)q_{jk}(v) && \text{if} \quad u \geq 1 \, . \end{aligned}$$

For the case $i = k$ we have

$$\begin{aligned} p_{ii}(u) &= 1 && \text{if} \quad u = 0 \, , \\ &= h_i(u) + \sum_{v=1}^{u} \sum_{j=1}^{N} p_{ij}(u-v)q_{ji}(v) \, , && \text{if} \quad u \geq 1 \, . \end{aligned}$$

Now let $H(u)$ be an $N \times N$ matrix with off-diagonal elements equal to zero, and i-th diagonal element equal to $h_i(u)$.

Also let $P(u)$ and $Q(u)$ be $N \times N$ matrices with (j,i)-th elements equal to $p_{ji}(u)$ and $q_{ji}(u)$ respectively. Then the above equations can be written in the matrix form

$$(3) \qquad P(u) = H(u) + \sum_{v=0}^{u} P(u - v)Q(v), \quad u \geq 0 \, .$$

Since $Q(u)$ contains the basic data of the semi-Markov model, and since $H(u)$ is calculated from this data using (iv), the longitudinal model matrices $P(u)$ are completely determined by solving (3).

Example 9: Continuation of example 8.

For the student example the values of $p_{ij}(u)$ for $i = 1,2,3,4,5,$ and $u = 0,1,2,\ldots,9$ are given by (to 2 significant figures)

i \ u	0	1	2	3	4	5	6	7	8	9
1	1	0.20								
2		0.65	0.23	0.05	0.01					
3			0.46	0.18	0.05	0.01				
4				0.34	0.14	0.02				
5					0.31	0.13	0.04	0.01		

Blank entries represent zero's or numbers less than .005.

Problem 7: Based on example 9 above.

a) Given that an individual enters as a freshman, what is the probability of graduation.

b) Given that the entering freshman eventually graduates, what are the mean and variance of the number of years spent as a student?

c) Given that the entering freshman drops out, what is the mean and variance of the number of years spent as a student? ■

If all the basic data (the $q_{ij}(u)$'s) are known, equation (3) shows that the longitudinal model matrices $P(u)$ can be calculated and all the results of Chapter III follow. Often the detailed transition probabilities are not known, and only estimates of the means and variances μ_{ij} and σ_{ij}^2 can be obtained, together with the q_{ij}'s. Even with this limited data it is often possible to obtain approximate results for the equilibrium behavior of the system.

Recall that $L = \sum_{u=0}^{\infty} P(u)$, and let $H = \sum_{u=0}^{\infty} H(u)$, and $Q = \sum_{u=0}^{\infty} Q(u)$. The equations in (3) can be written out as

$$\begin{aligned}
P(0) &= H(0) \\
P(1) &= H(1) + P(0)Q(1) \\
P(2) &= H(2) + P(1)Q(1) + P(0)Q(2) \\
P(3) &= H(3) + P(2)Q(1) + P(1)Q(2) + P(0)Q(3) \\
&\vdots \\
\text{etc}&.
\end{aligned}$$

Summing these equations and using the above definitions we get

$$L = H + LQ ,$$

or

$$L = H(I-Q)^{-1} .$$

Now H is the sum of diagonal matrices and is itself a diagonal matrix with (i,i)-th element equal to μ_i (see problem 5). Let $D = (I-Q)^{-1}$. Then d_{ik} is the expected number of visits to state i given that the system was entered in state k. Thus

$$\ell_{ik} = \mu_i d_{ik} ,$$

where ℓ_{ik} is the expected number of periods spent in class i, given the system was entered on chain k. If a stationary vector g gives the chain inflows in each period the steady state stocks will be

$$s = Lg .$$

Example 10: Continuation of example 9.

For the data given in the student example,

$$D = \begin{bmatrix} 1.00 & & & & \\ 0.75 & 1.00 & & & \\ 0.63 & 0.84 & 1.00 & & \\ 0.50 & 0.67 & 0.80 & 1.00 & \\ 0.48 & 0.64 & 0.76 & 0.95 & 1.00 \end{bmatrix},$$

$$H = \begin{bmatrix} 1.20 & & & & \\ & 1.25 & & & \\ & & 1.10 & & \\ & & & 1.05 & \\ & & & & 1.00 \end{bmatrix},$$

and

$$L = \begin{bmatrix} 1.20 & & & & \\ 0.94 & 1.25 & & & \\ 0.69 & 0.92 & 1.10 & & \\ 0.53 & 0.71 & 0.84 & 1.05 & \\ 0.48 & 0.64 & 0.76 & 0.95 & 1.00 \end{bmatrix}.$$

Problem 8: Based on example 10.

Assume that you enter this student group as a junior,

a) how many periods do you expect to attend?

b) what is the probability that you will graduate?

Problem 9: Show that, given you enter class k, the probability of ever reaching class i is d_{ik}/d_{kk}.

□

To continue with the steady state approximations consider next the case where input flows are growing geometrically at rate $(\theta-1)$. Thus $g(t) = \theta^t g$ and from equation 7 in III.4 the stocks in period t (t large) are given by

$$s(t) = \theta^t L(\theta) g \ ,$$

where

$$L(\theta) = \sum_{u=0}^{\infty} \theta^{-u} P(u) \ .$$

Now let $\theta^{-1} = \delta$, and define

$$\tilde{P}(\delta) = \sum_{u=0}^{\infty} \delta^u P(u) \ ,$$

$$\tilde{H}(\delta) = \sum_{u-0}^{\infty} \delta^u H(u) \ ,$$

and

$$\tilde{Q}(\delta) = \sum_{u=0}^{\infty} \delta^u Q(u) \ .$$

By multiplying the u-th matrix equation in (3) by δ^u and summing over u we get

$$\tilde{P}(\delta) = \tilde{H}(\delta) + \tilde{P}(\delta)\tilde{Q}(\delta) \ .$$

Thus

$$\tilde{P}(\delta) = L(\theta)$$

$$= \tilde{H}(\delta)(I - Q(\delta))^{-1} \ .$$

For δ close to 1 the basic approximation formulas (Appendix 1) can be used for the elements of $\tilde{H}(\delta)$ and $\tilde{Q}(\delta)$. From these

$$q_{ji}(\delta) = q_{ji}\delta^{\mu_{ji}}(1 + \frac{\alpha^2}{2}\sigma_{ji}^2)$$

and

$$\tilde{h}_i(\delta) = [1 - \delta^{\mu_i}(1 + \frac{\alpha^2}{2}\sigma_i^2)]/(1 - \delta) \ ,$$

where

$$\alpha = \log_e \theta, \ \theta = 1/\delta \ .$$

Example 11: Continuation of example 10.

Let $\theta = 1.03$, so that $\delta = 0.97$. Then

$$D(\delta) = \begin{bmatrix} 1.00 & & & & \\ 0.72 & 1.00 & & & \\ 0.59 & 0.81 & 1.00 & & \\ 0.45 & 0.63 & 0.77 & 1.00 & \\ 0.42 & 0.58 & 0.71 & 0.92 & 1.00 \end{bmatrix}$$

$$H(\delta) = \begin{bmatrix} 1.19 & & & & \\ & 1.24 & & & \\ & & 1.10 & & \\ & & & 1.05 & \\ & & & & 1.00 \end{bmatrix}$$

and

$$P(\delta) = \begin{bmatrix} 1.19 & & & & \\ 0.90 & 1.24 & & & \\ 0.64 & 0.88 & 1.10 & & \\ 0.48 & 0.66 & 0.81 & 1.05 & \\ 0.42 & 0.58 & 0.71 & 0.92 & 1.00 \end{bmatrix} .$$

The actual values of $P(\delta)$ are very close to these approximations. ■

5. A Theoretical Comparison.

The stochastic interpretations of the longitudinal and cross-sectional models developed in III.5 and II.1 are used in this section in a theoretical comparison of the two models. Some data on student enrollment is used to illustrate the results.

Throughout this section we assume the longitudinal model is a valid description of the system's law of motion. Our intention is to construct a good cross-sectional approximation to that model and then examine the quality of the approximation. The actual approximation is time dependent and also depends on past inflows. Moreover, it depends on data that is usually not available in a longitudinal model. Nevertheless, the approximation does help us to describe the rational limits of approximating a longitudinal model with a cross-sectional model.

Recall that $S(t)$ is an N-dimensional random vector, where $S_i(t)$ is a random variable which gives the stocks in class i at time t. The expected stocks in each class are given by the elements of $s(t) = E[S(t)]$. For a (possibly nonstationary) cross-sectional model the conditional expected value of $S(t + 1)$ given both the realized values of stocks $S(t)$ at time t and the (expected) inflows $f_0(t + 1)$ in period $t + 1$, is easily derived from equation (2) in II.2. Let the superscript c represent the "cross-sectional model.' Then

$$(4) \qquad E^c[S(t + 1)|S(t) = x] = Q(t)x + f_0(t + 1) .$$

Note that we use $Q(t)$ to indicate that the transition matrix can be non-stationary from period to period.

The basic longitudinal model gives the unconditional expected values of $S(t + 1)$. From equation (4) in III.2.

$$E[S(t + 1)] = \sum_{u=0}^{\infty} P(u)g(t + 1 - u) . \tag{5}$$

To compare the longitudinal and cross-sectional models we must derive an expression for the conditional expectation $E^{\ell}[S(t + 1)|S(t) = x]$, where the superscript ℓ denotes "longitudinal model." In order to determine this expression some assumptions must be made on individual behavior and some results of probability theory exploited.

The longitudinal model stipulates that each individual in the system is subject to a stochastic law of motion that depends only on the individual's chain and elapsed time in the system. In particular, the movement of any given individual is independent of the movement of others.

With each individual who enters the system we associate a counting random variable. Let

$$\begin{aligned} Z_{i,k}^{(j)}(t - u,t) &= 1 \quad \text{if individual } j, \text{ who entered in chain } k \text{ in period } t - u \\ &\qquad \text{is in class } i \text{ at time } t, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Recall that $g_k(u)$ is the total number who enter in chain k in period u. Then the stock in class i at time t is the random variable

$$S_i(t) = \sum_{k=1}^{K} \sum_{u=0}^{\infty} \sum_{j=1}^{g_k(t-u)} Z_{i,k}^{(j)} (t - u,t) . \tag{6}$$

The central limit theorem of probability theory states that under our assumptions $S_i(t)$ has approximately a normal distribution. Also the elements of the N-vector $S(t)$ are jointly normally distributed, and the elements of the 2N-vector $(S(t),S(t + 1))$ are also jointly normally distributed.

Now let $b_{ij} = \text{Cov}[S_i(t),S_j(t)]$, where Cov indicates covariance. Also let $c_{ij} = \text{Cov}[S_j(t),S_i(t + 1)]$. The matrices B and C, with (i,j)-th elements equal to b_{ij} and c_{ij} respectively, are N × N covariance matrices. From the theory of multivariate normal distributions we can now write down the expression for the conditional expectation, namely,

$$(7)\quad E^{\ell}[S(t+1)|S(t)=x] = C(t)B^{-1}(t)x + P(0)g(t+1) + [s(t+1)-(C(t)B^{-1}(t)s(t)+P(0)g(t+1))].$$

This complicated expression reduces to

$$E^{\ell}[S(t+1)|S(t) = x] = s(t+1) + C(t)B^{-1}(t)[x - s(t)] ,$$

so that when $x = s(t)$, the forecast reduces to $s(t+1)$.

Before we can compare the forecasts obtained in (4) and (7) it is necessary to analyse the covariance matrices $B(t)$ and $C(t)$. First consider $B(t)$. Using the expression in (6) with the definition of covariance one can show that

$$b_{ii}(t) = s_i(t) - \sum_{u=0}^{\infty} \sum_{k=1}^{K} p_{ik}^2(u) g_k(t-u) ,$$

$$b_{ij}(t) = - \sum_{u=0}^{\infty} \sum_{k=1}^{K} p_{jk}(u) p_{ik}(u) g_k(t-u) , \quad \text{for} \quad i \neq j .$$

Now let $M(t)$ be an $N \times N$ matrix with off-diagonal elements equal to 0 and $m_{ii}(t) = s_i(t)$. Let $G(t-u)$ be a similar $K \times K$ matrix but with $g_{ii}(t-u) = g_i(t-u)$. Then the matrix $B(t)$ can be written as

$$(8)\qquad B(t) = M(t) - \sum_{u=0}^{\infty} P(u)G(t-u)P'(u) .$$

Recall that the prime indicates matrix transposition.

We now turn to analyzing the matrix $C(t)$. Since $c_{ij}(t)$ is a covariance term between stocks in class i at time t and stocks in class j at $t+1$ it is necessary to know the joint distribution of the class of an individual at both t and $t+1$.

Define

$$f_{ij}^{k}(u) = \text{Prob}\left\{\begin{matrix}\text{In class } i \text{ at } t \ \underline{\text{and}} \\ \text{in class } j \text{ at } t+1\end{matrix} \middle| \begin{matrix}\text{entered chain } k \text{ in} \\ \text{period } t+1-u\end{matrix}\right\}.$$

Later in this section these joint probabilities are discussed in detail and related to results in III.10. Continuing with our analysis of $C(t)$ it follows from this definition of $f_{ij}^{k}(u)$ that, if $f_{ij}(t + 1)$ is the expected flow from class i to class j in period t + 1,

$$f_{ij}(t + 1) = \sum_{u=0}^{\infty} \sum_{k=1}^{K} f_{ij}^{k}(u + 1)g_k(t - u) .$$

Using (6) and the definition of covariance it can be shown that

$$(9) \qquad C(t) = F'(t + 1) - \sum_{u=0}^{\infty} P'(u + 1)G(t - u)P(u) ,$$

where F(t+1) is the N × N matrix of expected flows $[f_{ij}(t + 1)]$. Next, recall that $q_{ji}(t)$ is the fracti of those individuals in class i at time t who move to class j at t + 1. Thus

$$(10) \qquad q_{ji}(t) = f_{ij}(t + 1)/s_i(t) ,$$

or in matrix form,

$$(10) \qquad Q(t) = F'(t + 1)M^{-1}(t) .$$

Now clearly the stocks in class j at time t + 1 are given by the sum of all flows into class j in period t + 1. Thus

$$s_j(t + 1) = \sum_{i=1}^{N} f_{ij}(t + 1) + f_{0j}(t + 1) .$$

Using (10) and substituting for the input chain flows,

$$s_j(t + 1) = \sum_{i=1}^{N} q_{ji}(t)s_i(t) + \sum_{k=1}^{K} p_{jk}(0)g_k(t + 1) .$$

In matrix form this becomes

$$(11) \qquad s(t + 1) = Q(t)s(t) + P(0)g(t + 1) .$$

Equation (11) could have been obtained from (4) directly, but by fallacious reasoning. Recall that our assumption is that the longitudinal model truly describes movement through the system, whereas (4) is simply a cross-sectional representation which approximates the true model.

By subtracting (7) from (4) and substituting (11) one finds that

$$E^{c}[S(t+1)|S(t)=x] - E^{\ell}[S(t+1)|S(t)=x] = [C(t)B^{-1}(t)-Q(t)](s(t)-x) \ . \tag{12}$$

Equation (12) gives the one-period forecasting error caused by using the cross-section model in place of the longitudinal model. By taking expectations on S(t) we see that "on the average" the expected error is zero in every class.

In order to say more about the size of the discrepancy between the two models it is necessary to know something about the magnitude of the entries in the matrix $[C(t)B^{-1}(t) - Q(t)]$. Let

$$D(t) = \sum_{u=0}^{\infty} P'(u+1)G(t-u)P(u)$$

and

$$H(t) = \sum_{u=0}^{\infty} P(u)G(t-u)P'(u) \ .$$

Then from (8) and (9) we have

$$B(t) = M(t) - H(t)$$

and

$$C(t) = F'(t+1) - D(t) \ .$$

From these equations together with (10) it can be shown that

$$C(t)B^{-1}(t) - Q(t) = [Q(t)H(t) - D(t)]B^{-1}(t) \ . \tag{13}$$

Problem 10:

a) Verify equation (13).

B) Show that if $P(u + 1) = Q(t)P(u)$ for all $u \geq 0$, then $C(t)B^{-1}(t) - Q(t) = 0$ and the two models coincide. ■

To investigate (13) further we consider the one class, one chain model with constant input. In this case all matrices and vectors reduce to scalars, $g(t) = g$ for all t, and $P(u) = p(u)$. Moreover

$$H = g \sum_{u=0}^{\infty} p(u)^2 \ , \quad s = M = g \sum_{u=0}^{\infty} p(u) \ ,$$

$$F = g \sum_{u=1}^{\infty} p(u) \quad \text{and} \quad D = g \sum_{u=0}^{\infty} p(u)p(u + 1) \ .$$

Let $\lambda = \sum_{u=0}^{\infty} p(u)$, the expected lifetime of an individual in the system. Then

(14) $$QH - D = \frac{g}{\lambda}\left[\sum_{u\geq 0} p(u)^2 \sum_{u\geq 0} p(u + 1) - \sum_{u\geq 0} p(u + 1)p(u) \sum_{u\geq 0} p(u)\right] .$$

The term in parenthesis in (14) is

$$\sum_{u\geq 0} p(u)^2(\lambda - 1) - \lambda \sum_{u\geq 0} p(u)p(u + 1) = \lambda \sum_{u\geq 0} \Delta(u + 1)p(u) - \sum_{u\geq 0} p(u)^2 \ ,$$

where $\Delta(u + 1) = p(u) - p(u + 1)$.

Interpreting $p(u)$ as the tail distribution of a non-negative random variable, say Λ for "lifetime," one can show that

(15) $$\sum_{u\geq 0} p(u)[1 - p(u)] = \sum_{u\geq 0} \Delta(u) \sum_{v\geq u} p(v) \ ,$$

and

(16) $$\sum_{u\geq 0} [\Delta(u) + \Delta(u + 1)]p(u) = 1 \ .$$

Using (15) and (16) in (14) gives

(17) $$QH - D = \frac{g}{\lambda} \sum_{u\geq 0} \Delta(u)\left[\sum_{v\geq u} p(v) - (\lambda)p(u)\right] .$$

Let us assume now that the expected remaining lifetime of a person whose time in the system exceeds u time periods is no more than the expected lifetime λ of a new input. We say that people have "mean residual life" bounded above by the original mean life, and say that Λ has MRLA if

$$\sum_{v \geq u} \frac{p(v)}{p(u)} \leq \lambda, \quad \text{all} \quad u = 0,1,2,\ldots \quad \text{for which} \quad p(u) > 0 \ .$$

Note that equality holds in this equation for the geometric distribution. Table IV.1 shows that in a particular case of students attending the University of California at Berkeley, (see Table II.15 also) this assumption is valid.

Under the MRLA assumption, from (17) we see that

$$QH - D \leq 0 \ .$$

In the stationary case $[QH - D]B^{-1}[s - x]$ is independent of t.

Since B^{-1} is nonnegative, we have the following conclusions:

If we assume Λ has MRLA,

a) If $x < s$, the cross-sectional model under-estimates the value of $E^{\ell}[S(t + 1)|S(t) = x]$.

b) If $x > s$, the cross-sectional model over-estimates the value of $E^{\ell}[S(t + 1)|S(t) = x]$.

Since $S(t)$ has a marginal normal distribution we can say more about the expected error in the one dimensional case. The error is a normal random variable with zero mean, and variance equal to $(QH-D)^2B^{-1}$ (where these are all scalars). Thus we can say that with probability about .95 the error will lie in the interval $(-2B^{-1/2}|QH-D|, + 2B^{-1/2}|QH-D|)$. The length of this interval increases as the square root of g. However, s the expected value of $S(t)$ increases as g. Thus the interval length divided by s, or the fractional error range, decreases

Lifetime (semesters) u	$\Pr[\Lambda>u] = p(u)$	$\sum\limits_{v\geq u} p(u)$	$\sum\limits_{v\geq u} p(u)/p(v)$
0	1.000	6.959	6.96
1	0.972	5.959	6.14
2	0.905	4.987	5.52
3	0.756	4.082	5.42
4	0.684	3.326	4.86
5	0.593	2.642	4.47
6	0.562	2.049	3.65
7	0.524	1.487	2.84
8	0.498	.936	1.88
9	0.199	.465	2.34
10	0.130	.266	2.05
11	0.050	.136	2.72
12	0.036	.086	2.39
13	0.017	.050	2.94
14	0.015	.033	2.20
15	0.011	.018	1.64
16	0.007	.007	1.00

Table IV.1. Mean Residual Life of Freshman Students Entering U.C. Berkeley in Fall Semester, 1955.

as the square root of g. So as g increases, and hence s increases, the width of the confidence interval of error increases much more slowly. To illustrate this we use the lifetime distribution from Table IV.1, and for various cohort sizes we show how the interval length changes. The results are given in Table IV.2. It is clear from this table that even though the lifetime distribution differs considerably from a Markovian (geometric) distribution with the same mean, the confidence intervals on the forecasting error are extremely small relative to the expected number in the system. For comparison $p(u)$ is drawn in Figure IV.1 together with a geometric distribution.

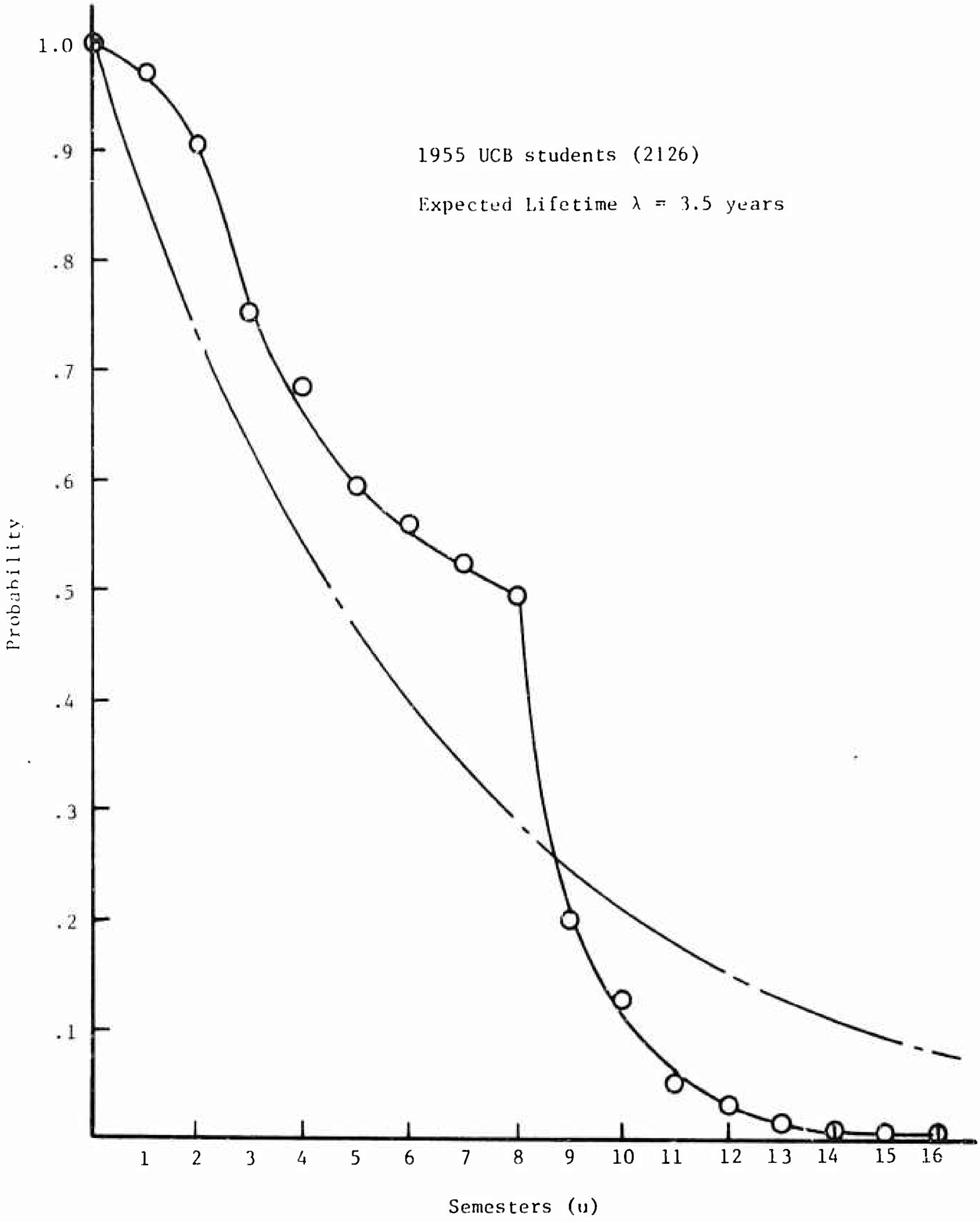

Figure IV.1. Comparison of p(u) for UCB Students with a Geometric Distribution.

Cohort Size g	E[S] = s	Confidence Interval for Forecast error
1000	6,959	(-7,7)
2000	13,918	(-10,10)
3000	20,877	(-12,12)
4000	27,836	(-14,14)

Table IV.2. 95% Confidence Intervals for Various Cohort Sizes.

Determination of properties of the matrix in (13) for the multi-class, multi-chain case is much more difficult than in the one-class, one-chain case. A 4-class, 4-chain numerical example is given which uses the student enrollment data from Table III.3, and assuming constant cohort size input.

The forecasting error given by (12) has a multivariate normal distribution with mean 0 and covariance matrix $(QH-D)(B^{-1})'(QH-D)'$. Using the data given in Table III.3 for freshmen, sophomores, junior and seniors at the University of California, Berkeley 1955-1969, calculations were made assuming constant cohort sizes of 3000 freshmen, 700 sophomores, 1300 juniors and 150 seniors entering each fall semester. These figures are approximately what the Berkeley campus had been experiencing in its fall new admissions.

Table IV.3 gives the matrix B, whose $(j,i)^{th}$ element is the covariance of $S_i(t)$ and $S_j(t)$ for some t. Also included is s, the vector of expected stocks in each class.

Class j \ Class i	Freshmen	Sophomores	Juniors	Seniors
Freshmen	673	-454	-30	-10
Sophomores	-454	1453	-380	-43
Juniors	-30	-380	2137	-535
Seniors	-10	-43	-535	2216
Expected Values	3868	3324	4687	3227

Table IV.3. Covariance Matrix B for the 4-class example.

The variance of the number in each class increases as the class increases, and all classes are negatively correlated.

Table IV.4 gives the matrix $(QH-F)B^{-1}(QH-F)'$, which is the covariance matrix of the forecasting error. It can be seen that these numbers are very small compared to the size of the predicted values, as was found in the single state case.

Class j \ Class i	Freshmen	Sophomores	Juniors	Seniors
Freshmen	6.7	2.2	-22.4	-5.4
Sophomores	2.2	1.0	-8.5	-2.7
Juniors	-22.4	-8.5	82.2	29.5
Seniors	-5.4	-2.7	29.5	41.8

Table IV.4. Covariance Matrix of Forecasting Error.

The matrix $(QH-D)B^{-1}$ is given in Table IV.5.

Class i Class j	Freshmen	Sophomores	Juniors	Seniors
Freshmen	.068	.013	.002	.001
Sophomores	-.041	-.003	.003	.001
Juniors	.290	-.062	-.030	.029
Seniors	.040	-.046	-.125	.032

Table IV.5. $(QH-D)B^{-1}$ for the 4-Class Example.

This is an example where $(QH-D)$ is neither $\geq$ nor ≤ 0, unlike the one-class, one-chain model.

Even though movement through the system is far from that represented by a stationary cross-section model (i.e., $P(u) \neq Q^u$ for some Q), when constant cohort sizes are used the cross-sectional model gives essentially the same prediction as the more complex cross-sectional model. However, the longitudinal model is primarily formulated for forecasting under conditions of controlled input. This is often the situation when policy changes are implemented, and under such conditions the sizes of cohorts is successive time periods can and do vary considerably. For example, the freshmen cohorts in the fall quarters at Berkeley in the period 1966-1969 are shown in Table IV.6. This was a period when total campus enrollment was controlled, and new students entered only to fill available room.

Date	Cohort Size
Fall 1966	3,053
Fall 1967	3,303
Fall 1968	2,239
Fall 1969	1,883

Table IV.6. Freshmen Cohort Sizes at U.C. Berkeley

One can see that, since $F(t)$ and $s(t)$ are both functions of previous cohort sizes (up to period t), that the cross-sectional transition probabilities will change with time, and that estimating them from cross-sectional data in two consecutive years will not account for gross changes in cohort sizes.

We end this section with a brief discussion of the joint probabilities $f_{ij}(u)$ and their connection with the flow parameters $p_{ik}(u)$ in III.10 (longitudinal conservation).

First it is easy to see that if $u = 0$, then $f^k_{0j}(0) = p_{jk}(0)$, and for $i \neq 0$, $f^k_{ij}(0) = 0$. These relations follow directly from the definitions. Next, since any individual who leaves the system cannot return, for j and $u \geq 1$ $f^k_{0j}(u) = 0$. Also, by looking at the flows into some state j in period u it follows that

$$p_{jk}(u) = \sum_{i=0}^{N} f^k_{ij}(u)$$

Similarly, by looking at flows out of some state i in period $(u + 1)$,

$$p_{ik}(u) = \sum_{j=0}^{N} f^k_{ij}\ (u + 1).$$

Thus the $\{p_{ik}(u)\}$ are the marginals of the joint probabilities $\{f_{ij}(u)\}$. In many applications the $\{f_{ij}(u)\}$ are hard to measure and it would be advantageous if they could be estimated from the $\{p_{ik}(u)\}$ which are relatively easy to measure. In general the marginals do not determine the joint distributions.

Problem 11: The longitudinal model would have serial independence if $f_{ij}^{k}(u + 1) = p_{ik}(u)p_{jk}(u + 1)$. Since people who leave cannot return, $f_{0j}^{k}(u) = 0$ for all u. Use this to prove that we cannot have serial independence in the longitudinal model.

6. Notes and Comments.

The material in section 3 is based on Hayne [1974] and Hayne and Marshall [1974]. This type of model makes it possible to work with a highly disaggregated manpower classification scheme and still have some control over the interpretation and manipulation of the model.

The semi-Markov model of section 4 is new. The reader may consult Ross [1970], and references cited there, for a decription of semi-Markov models. Austin [1971] and Bartholomew [1973] discuss semi-Markov models. The treatment in section 4 is quite different. We stress approximations that can be obtained from the transition probabilities, and the first two moments of the length of a visit.

Section 5 is based on Marshall [1973]. It reveals the underlying structure of the longitudinal models and reinforces the theoretical notions derived in section 10 of chapter III.

SELECTED BIBLIOGRAPHY

Austin, R. W., "A Semi-Markov Model for Military Officer Personnel Systems," M.S. Thesis, U.S. Naval Postgraduate School, Monterey, CA (Sept. 1971).

Bartholomew, D.J., Stochastic Models for Social Processes, second edition. J. Wiley and Sons, New York (1973)

Hayne, W.J., "Analysis of Some Two-Characteristic Markov-Type Manpower Flow Models with Retraining Applications," Ph.D. Thesis, Naval Postgraduate School, Monterey, CA (June 1974).

Hayne, W.J., and Marshall, K.T., "Two-Characteristic Markov-Type Manpower Flow Models," Tech. Report NPS55Mt74071, Naval Postgraduate School, Monterey, CA (July 1974).

Marshall, K.T., "A Comparison of Two Personnel Prediction Models," Operations Research, Vol. 21, pp 810-822.

Ross, S.M., Applied Probability Models With Optimization Applications, Holden-Day, San Francisco (1970).